

Tracks Through Time and Continuous Processes: Transitions, Sequences, and Social Structure

Brendan Halpin
University of Limerick*

March 2003

1 The richness of longitudinal data

In the social and economic sciences it appears that there was a pioneering enthusiasm for longitudinal data in the late 1960s and early 1970s, resulting *inter alia* in the US Panel Study on Income Dynamics (1968) and life-history studies such as Natalie Rogoff Ramsøy's Norwegian Life History Study (Rogoff Ramsøy, 1975). While the continuity of the PSID helped a lot in the development of techniques for the use of individually longitudinal data, it has been a slow development, but has been nonetheless relentless and cumulative. By now we have a very substantial array of panel and life-history studies, longitudinal elements incorporated in exercises such as the Labour Force Survey, and a growing battery of longitudinal data sets based on official data collection. Commensurate with the rich data is a growing body of high-quality research that takes full advantage of its longitudinality. The argument no longer needs to be made for the greater richness, power and sheer interest of longitudinal data.

Technical development has also been slow, but has been cumulative and is by now very substantial.¹ 1984 saw sociologists become aware of what has become the *iconic* technique for the analysis of longitudinal data, certainly on the sociological side, hazard rate modelling or 'Event History

*Paper prepared for the Conference, 'Frontiers in Social and Economic Mobility', Cornell University, March 27–29 2003 (tracks.tex,v 1.4 2003/04/04 11:30:18 brendan Exp)

¹By development here I mean not simply the invention of new techniques, but also their diffusion into the imaginations of researchers, and researchers' overcoming of the often substantial practical difficulties of handling longitudinal data

Analysis', with the publication of Tuma and Hannan (1984) and Allison's Sage book. EHA is by no means the only important technique: loglinear based longitudinal modelling is common (*e.g.*, Hagenars, 1990), and 'cross-sectional time-series' (*e.g.*, Greene, 2003) and GEE approaches of largely econometric origin are also important. I mention these not as an exhaustive list but as three typical approaches, focussing on different aspects of the individual-level longitudinality: respectively duration in states, matrices of probabilities of transitions between states, and the co-evolution of quantities through time, particularly for those who work mainly with discrete state spaces. These techniques are extremely powerful for certain sorts of research question, more powerful perhaps the more specific the question. They also exploit different aspects of the data and are therefore complementary in many respects. However, individually they give an incomplete picture of longitudinal reality and there are bounds to the sorts of information they can provide us. In the present paper I want to focus on two sorts of shortcoming:

- the micro bias of individually longitudinal approaches, and
- the lack of an overview or holistic perspective.

For the former, I will focus mostly on EHA, which has special importance by virtue of its wonderful match with the way we think about causality operating through time. The latter gives me an opportunity to talk about sequence analysis techniques (such as the Optimal Matching algorithm), which are both an extremely useful addition to the longitudinal analyst's toolbox, and for the moment at least, subject to some very serious limitations. My opinion is that because of their limitations, holistic approaches such as sequence analysis have a largely exploratory value. I will finally offer a few thoughts on the role of the exploratory in theoretically informed empirical work.

2 EHA and micro bias

Event history analysis or hazard rate modelling is peculiarly attractive to the longitudinal analyst, because of the close mapping between the statistical model and a causal model that coincides well with our experience of "things happening through time". In the context of a discrete state space, the hazard model with fixed and time-dependent covariates fits very well with our lived experience ("How long will it take me to get promoted? How would it be different if I were a woman? How much quicker if I get my PhD? if I publish in the AJS?"). The attractiveness of the mapping between statistical and causal models is similar to that of regression analysis but with the large advantage of the incorporation of time in a very realis-

tic way. Raftery (2001) suggests regression is popular because it maps well onto how sociologists think about causality; Abbott (1988) expresses this more negatively as sociologists allowing their thinking to be straitjacketed by “General Linear Reality”. Raftery and Abbott agree that EHA is conceptually a descendent of regression and thus subject to the same claims, but I would be very reluctant to endorse Abbott’s criticism with reference to EHA because I view the causal model implied by EHA to be extremely powerful and verisimilar – time indeed has a direction and one’s future is determined (in a probabilistic way) by one’s history, attributes and present circumstances; we *ought* to think in these terms in our role as social scientists, as well as that of statistical analysts. Of course there are respects in which this model is inadequate – intentionality and planning being one major source of problems – and I will deal with this issue in section 3. In this section I want to deal with the problem of the micro bias that arises from this intensive use of rich individual level data.

2.1 Micro versus macro

Coleman (1962) famously admonished sociologists to avoid the error of analysing “not the social system but the IBM cards”, but the wave of individual level data that has showered us in the intervening four decades has tended to make us deaf to this particular entreaty. This is not necessarily qualitatively different for individually longitudinal data, but perhaps we are here in degree even a little deafer, simply because there is so much of interest to look at. I would distinguish between methodological individualism as a conceptual framework (a perspective I imagine many of the sociologists present endorse to a greater or lesser degree), and this *de facto* individualism: the former starts from the individual but has no difficulty in dealing with supra-individual structures that emerge, while the latter tends to have difficulties in dealing with supra-individual structures or simple aggregates. Here we do have a context in which allowing the technique to guide our theoretical thinking can have a limiting effect.²

Let me proceed by way of an example from my own experience (while offering the caveat that the problems reported may have as much to do with limitations in my statistical skills as that of the technique). The key problem in the following is that to use only individual attributes to predict an outcome that is also dependent on the social structure, will lead to models that are mis-specified and potentially misleading. However, to incorporate information about the social structure in the individual-level

²This issue is not entirely new: there is a tradition stretching back to Sørensen and Tuma (1981) *inter alia* which looks at mobility in more or less closed systems such as organisations, in terms of “vacancy chains” and similar concepts.

model is not necessarily easy³. Sometimes the solution is to move to a less sophisticated class of model, in the following case from EHA (which Raferty (2001) classes as a later development in the ‘second generation’ of statistical models in sociology) to loglinear analysis (‘first generation’ for Raftery).

The problem domain for my example is assortative mating: the extent to which there is systematic association between one’s education and that of one’s spouse. We can view this as an outcome or a social structure, or in respect of the dynamic processes in the mate ‘market’. A few years ago a colleague and I (Chan and Halpin, forthcoming) participated in a fascinating collaboration organised by Andreas Timm and Hans-Peter Blossfeld (Blossfeld and Timm, forthcoming). Blossfeld and Timm established a template for the research (an event history analysis of partnership formation, taking the educational level of the respondent and spouse into account (Blossfeld et al., forthcoming)) and assembled a team of country experts who applied to template to their own nations – a pragmatic and often highly effective structure for comparative research.

There is a lot to recommend an event-history approach to the issue of assortative mating. The structure of association between spouses’ education levels is the outcome of complex social processes, some of which we have access to in longitudinal data sets, and we can thus examine the whole process of partnership formation in a dynamic framework, and see precisely how the education of the individual affects his or her chances of marrying, controlling also for the educational level of the spouse (via a competing risks model with marriage to a spouse of each educational level as the competing outcomes). There are so many advantages to this approach – we take account of the timing of marriage; we use information from those whom we do not observe to marry as well as those who do; we simultaneously take account of other individual and temporal characteristics; we develop a model directly explicable in terms of real observable micro-social processes – that it is hard to see how reducing this rich data to a thin table of spouse pairs classified by their education levels could be better in any respect. Nonetheless, our experience left us unsatisfied until we had conducted a re-analysis (Halpin and Chan, 2002) of the data using loglinear analysis of simple tables of spouse pairs, following Mare (1991).

Why? Two reasons: first, one important question cannot be addressed properly in the framework – is the structure of assortative mating weaker or stronger over time or across countries? Secondly, unless we explicitly

³But not impossible: models of the hazard of exiting unemployment sometimes successfully incorporate time-series of local labour market statistics

take account of the structure of opportunity and competition (the educational distribution of competitors and of potential spouses) it can be argued that the parameter estimates of all variables in the model are subject to missing variable bias. These are really two faces of the same problem: the emergent structure of assortative mating arises from the interaction of individual action with supra-individual structures – individual partner search in the context of a changing educational distribution of competitors and potential spouses. That is, the social processes that operate, for instance a positive preference for homogamy, do so in the context of the distributions and interact with them to produce the outcome structure. Log-linear models of tables of spouse pairs deal with this very nicely, completely separating the distributional effects (the ‘margins’) from the net association, which can be considered the net effect of the social processes. Unfortunately, loglinear models are by definition data poor (*i.e.*, small numbers of variables) and they really go very little farther than this helpful separation of the structure from the process, leaving us with the net association as a black box.

The EHA approach at least in principle allows us to unpick the social processes, but it is very hard to “decontaminate” them, to remove the structural component. We attempted to deal with this by incorporating measures of the “opportunity structure”. We constructed a matrix of age and period specific educational distributions of single people of the opposite sex, and included this as a time-dependent covariate (see Figure 1). Though an unwieldy strategy, it had partial success: parameter estimates for other variables changed (we assume for the better!) and the structure terms themselves were generally very significant. However, their parameter estimates were difficult to interpret – for instance, the chance of marrying a bride with incomplete second level rises more in response to a rise in the proportion of single women with complete second level than of those in the appropriate category. This is obviously due both to the high collinearity of the proportions, and to the fact that the distribution of potential spouses is strongly associated with the distribution of competitors, a factor we did not enter in the modelling. We considered at that point entering some sort of cross-product of the opportunity and “competition” structures, but decided to revert to loglinear models rather than attempt to reinvent them. There were serious costs to this: we become blind to the timing of marriage in the life course, to those who do not marry, and have no substantive covariates; however, we immediately solve the problem of separating structure from net association.

In summary, we face a trade-off: EHA’s dynamic individual focus is in many ways richer, certainly in terms of the causal and sequential nar-

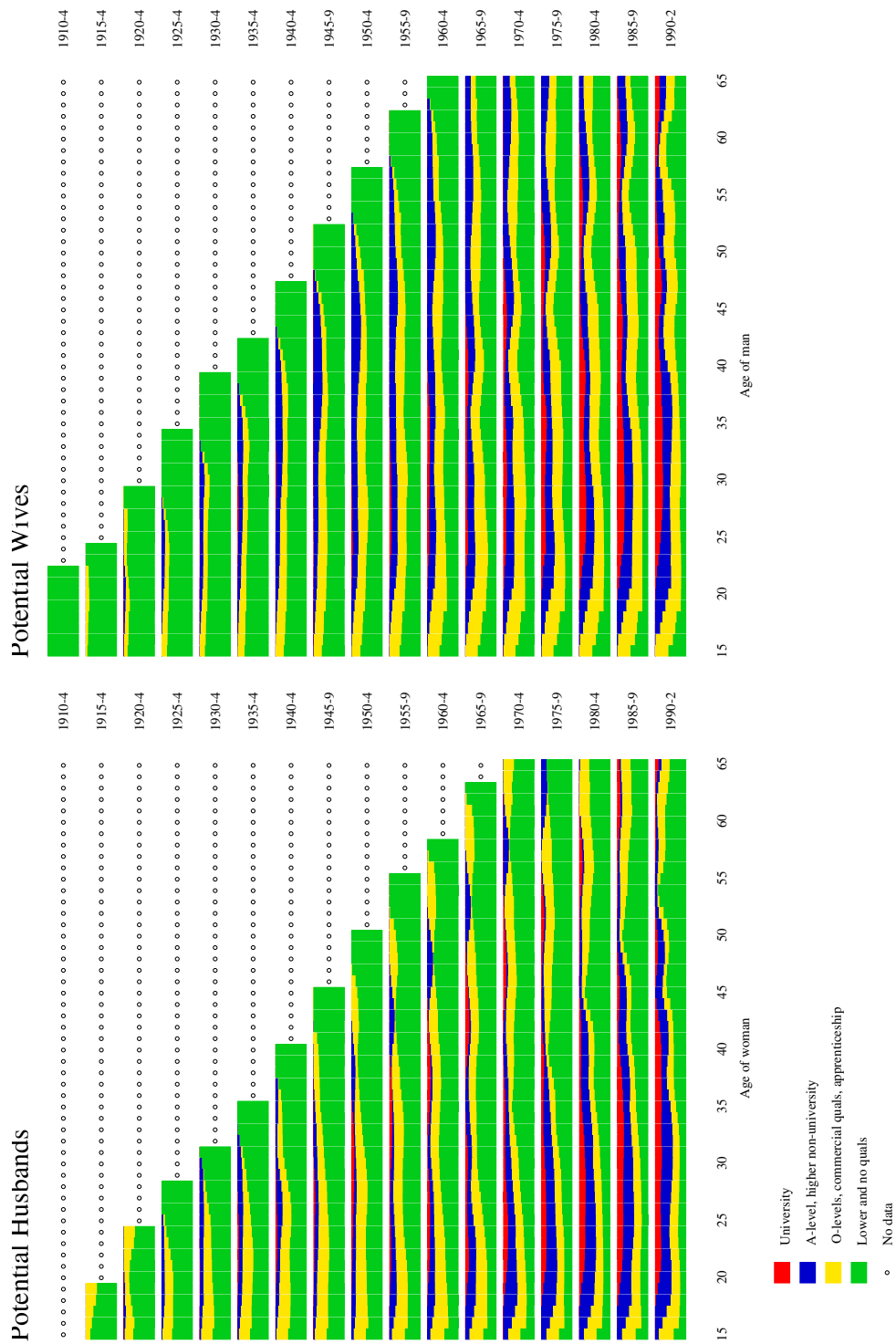


Figure 1: The “Opportunity Structure”. For each combination of calendar year and year of age, this represents the relative distribution of education of single people in a five-year band around the respondent’s age, offset so that females are on average two years younger than their spouse.

rative it allows, but it fails to deal with structure, in a way that *may* be catastrophic. On the other hand, loglinear models succeed with ease in extracting net association but give us no information whatsoever about what brings it about.

3 Holism

There is a second class of ways in which the story we get from an EHA perspective, or from other perspectives that track change through time in a similar manner, may be inadequate. If the pattern of transition probabilities that the models estimate is in some respect sufficiently complex, it may not be adequately represented by a simple EHA model. This complexity may come in many forms. Perhaps the effect of covariates changes with calendar time⁴, or with age⁵, or in some other unexpected respect. Or perhaps humans can warp the arrow of time such that the future, or more properly expectations about the future, affects the present.⁶ Conventional or normative structures may also be important – ‘finish education first, then marry’, ‘marry first, then have children’, particularly if multiple related norms have interactive effects. In such cases, it is possible that the longer term trajectory might differ significantly from what might be projected from the results of an overly simple EHA model, or models of annual transition rates.

To some degree we can cope with this by using a variety of measures, such as making pairwise comparisons of status at widely separated parts of the career (*e.g.*, social class at entry to the labour market, and social class at age 35) as well as annual transition rates or instantaneous rates of the hazard of inter-class moves in EHA models. These approaches, each of which has a very narrow focus, are quite complementary. Nonetheless, it is reasonable to seek techniques which treat the trajectory as a whole, and this accounts for the enthusiasm with which sequence analysis (SA) methods such as the Optimal Matching Algorithm (OMA) are being taken up in a wide variety of social research contexts (two recent examples I have come across are Scherer (2001) and McVicar and Anyadike-Danes (2002), both in the substantive area of early labour market careers, where OMA seems to have caught the imagination).

⁴Due, perhaps, to structural effects such as those discussed in section 2.

⁵Perhaps, as Billari (2002) points out, to life-course considerations such as ‘permanent income’.

⁶Jill Pearson, a PhD student in UL, adds to ‘career ladder’ the notion of the ‘career snake’: the lateral and downward move made to escape a career dead-end, in the expectation of future advancement.

There is another reason for wanting to analyse the whole trajectory, even when it displays no such complexity: it can simply be more interesting, easy to interpret and easy to report, than the results of painstaking modelling. It is all very well to know that we have certain patterns of transition probabilities, and that they are affected in certain ways by certain other factors, but it is quite another thing to imagine what the resulting distribution of career types might be.

However, I will stress this point: *trajectories are epi-phenomena, but the operation through time of the underlying transition probability structure is fundamental*, however complex that may be. Our goal should always be to understand the latter, but under some circumstances the holistic focus on the trajectory will help.

3.1 Before OMA

The optimal matching algorithm was not a solution in search of a problem, as attest the many attempts to develop classifications of sequences before OMA emerged. Particularly when one is thinking of careers or trajectories that might be in some sense problematic (for instance, the transition from school to the labour market, particularly in poor economic circumstances) or subject to rapid change (*e.g.*, the transition from family of origin to adult- or parenthood), the idea of a technique that can generate an empirical typology of trajectories is very attractive. Ad hoc approaches include Degenne et al. (1996), who divided school-to-work careers into six month segments, made a number of summaries of the segments, followed by a factor analysis of the summaries, and the development of a typology by cluster analysis. Roundabout, but effective and largely comprehensible. Buchmann and Sacchi (1995) reduced a large state space (occupations) to a small one by factoring and clustering, and then created inter-sequence differences by calculating period-by-period distances between sequence pairs. Again comprehensible, but it suffers from the serious problem that sequences cannot be aligned: month 1 is compared with month 1, month 2 with month 2 and so on, so similar sequences which are a little out of phase may appear very different. More sophistication is evident in Dijkstra and Taris (1995), particularly in that they attempt to generate a formal methodology for comparing sequences, though one subject to a number of sustained criticisms (*e.g.*, Abbott, 1995).

3.2 Optimal Matching

Andrew Abbott, with the assistance of a cast of dozens of co-authors, is largely responsible for the popularity of OMA in sociology, a veritable evangelist. In large measure his enthusiasm is justified, because for the

task of calculating inter-sequence similarity scores OMA is clearly better, more general and more flexible than anything which has gone before. Insofar as he seems to suggest that OMA ushers in a new era of sociological research, I am afraid I part company with him! I am happier with Raftery's 2001 characterisation of OMA as one of a number of "third generation" techniques which are attempting to go beyond the linear-regression family and come to grips with larger structures such as networks, macro-social structures or trajectories⁷.

This is not the place to rehearse the detail of the optimal matching algorithm, and I hope the following will suffice: OMA takes pairs of sequences and calculates the cost of changing one into the other by a series of "elementary operations" of insertion, deletion and substitution. The sequences may be anything: often biological macro-molecules (the biggest application is with analysis of DNA and proteins), computer files, bird-song (*e.g.*, Bradley and Bradley, 1983), dance patterns (Abbott and Forrest, 1986), the rhetorical structure of journal articles (Abbott and Barman, 1997), the school-to-work transition (McVicar and Anyadike-Danes, 2002; Scherer, 2001) or class careers (Halpin and Chan, 1998). For some of these examples I withdraw my claim above that the trajectory or sequence is an epiphenomenon: molecules, songs and dances, and journal articles have a sequential structure that is a thing-in-itself, and which only incidentally, if at all, unfolds through time; the application of OMA to such sequences therefore has a much sounder footing. However, most sequences of interest to social scientists (and probably all relevant to social mobility) do develop stochastically through time, where the facts of the earlier parts determine the probabilities of the later. That is, for most sequences we will rightly be more interested in the processes that created them than their actual track through time.

3.2.1 CRITICISMS OF OMA

OMA is somewhat controversial, promoted with enthusiasm through the late 1980s and 1990s by Abbott, but attracting criticism on a wide range of fronts, for instance in the debate in *Sociological Methods and Research* (Abbott and Tsay, 2000; Levine, 2000; Wu, 2000; Abbott, 2000). I find myself ambivalent with regard to many of the criticisms advanced. This is because they may reflect higher expectations of SA than I have – I feel it is justified by, but limited to, its exploratory role, at least for now, while many of the critics judge it against the standards of analytical statistical

⁷I agree with him in particular that "a more explicitly model-based approach would help" with OMA (Raferty, 2001, section 4.3).

techniques. I agree with the substance of several of the criticisms but find many of the problems tolerable in an exploratory context. For instance, if minor changes to the input give different results in a formal hypothesis-testing model, one would be right to be worried about the robustness of the finding, but with an exploratory technique one should simply be tempted to change the settings again to explore what else might happen: that is, with exploratory techniques we are not looking for a 'right answer', we are simply looking.

In particular, that would be my reaction to some of the worries expressed about the determination of the values in the substitution matrix, which feeds into the calculation of the intersequence distance in a critical way. Finding the 'right' values may be a worthwhile task, and for certain domains and certain problems experience may allow us to tune the substitution matrix (as seems to be the experience in molecular biology, but they have huge experience and a very narrowly defined task), but worries that we have no theoretical basis to determine the substitution matrix, and therefore no way to justify our analysis seem to me to be alleviated by regarding the exercise as limited to the exploratory. I feel the answer is to "suck it and see": examine your state space, and make an assessment of the difference between the states, from any reasonable point of view. The algorithm simply translates that structure of the state space into distances between sequences in the state space. If you don't like the results, change the matrix and go again!⁸ If we regard the exercise as exploratory, we have no guilt about fishing for significant results. "If we define the state space as having this structure, OMA shows us this structure in the trajectories."

Another criticism with which I have limited agreement is that the elementary operations of insertion, deletion and substitution have no clear sociological interpretation. With the exception of their consequences for the issue of duration (on which more below), I feel this is largely a red herring: these steps are simply a way of calculating a distance score between sequences, one that is computationally efficient. By analogy, we need not worry about the lack of a sociological interpretation of the Newton-Raphson method in maximum-likelihood estimation. This criticism is sometimes bolstered by pointing out that there is a better match with the biological generation of sequences. This is only partly true: biological sequences are "cut and pasted" but in fact the match between that and the elementary operations is not that close.

⁸This *will* change the outcome: I recently ran some tests using TDA's (Rohwer, 1998) optimal matching commands, and cross-tabulated cluster membership given different substitution matrices – there tends to be association but it is far from perfect.

Wu (2000) has formulated a number of other criticisms. One is that the elementary operations are blind to the arrow of time. It is completely unnecessary for them to consider the arrow of time because this information is encoded in the sequences – we get meaningful results because there is meaning in the sequences.

Related is the worry about symmetry: that substitution matrices cannot be asymmetric – in so far as this is based on the notion that transition probabilities may be asymmetric, I will note that there is no requirement that substitution costs should be related to transition probabilities because substitution *does not represent transition* (however, transition probabilities may be a perfectly good way of assessing similarities in the state space). Furthermore, asymmetric substitution costs imply asymmetric distances between sequences, which involves a complication way over and above refinements of the OM algorithm.

There is a logically distinct argument about symmetry, which boils down to the equality of the distance between subsequences pairs such as

- ABC and
- ABD

and

- CBA and
- DBA.

I can see that this could in certain circumstances be by an unwelcome restriction, but in most cases the logic that makes the distinction important will also mean that these subsequences will be embedded in quite different contexts, and therefore empirically differentiated.

The issue of duration is where I have my strongest worries about the applicability of OMA to the sort of data I work with: life histories of one sort or another (class careers, for instance) where duration is important and the average duration is well above 1 unit. We therefore represent trajectories as strings with large elements of repetition, *e.g.*, AAAAAAAAAA ADDDDDEEEE EEEEEEEEEEE EAAAAAAAAA, where the amount of repetition indicates the duration in each state. Since the elementary operations are literally element-wise, they make no distinction between deleting a B in a sequence of ten and deleting a it in ABC. This means sequences like ABC and ABBBBBC will be considered as relatively more different than they are in my opinion as an analyst. There is a limited solution to this in TDA (Rohwer, 1998), where it is possible to specify a cost function, $\gamma = \alpha + \beta \times \text{indel}$ to weigh insertions and deletions more lightly if they are expanding or reducing a contiguous run of the same state. However, I believe this to be at the rather high cost of constraining substitution costs to be constant.

3.2.2 SOME INNOVATIONS

The field is developing apace, and while OMA is top dog now, with a very useful implementation in TDA, new things are coming. Bayesian approaches such as Gibbs Sampling (*e.g.*, Abbott and Barman, 1997) and Markov Chain Monte Carlo models (both enumerated in Raftery's 2001 third generation) seem to offer improvements in both generality and specificity. Elzinga (2003) has proposed an alternative algorithm to OMA, which may supersede it, possibly having advantages concerning duration, and the importance of the order of non-adjacent states. The algorithm consists in an efficient way of counting the number of common n -tuples in pairs of sequences (ABC and ABDC have in common AB, AC, BC and ABC). For those who have reservations about OMA's elementary operations, this may offer some relief (though in practice its results seem not to be radically different).

Billari et al. (2000) are an example of a completely different approach. They apply machine learning techniques (decision trees and rule generators) to data on life courses (coded in terms of whether a given event had happened, its timing and its order relative to other events – non-repeatable events such as first leaving home, leaving education, forming a partnership, marrying, having a child). In their example the sequence or order information (more than timing or status) served to discriminate well between Austrian and Italian samples of young adults' careers. The advantage of the technique is the interpretability of the rule sets or decision trees that emerge, along with some measures of how important certain distinctions are. We can see a relationship with the Elzinga technique, with the focus on ordered pairs (but not of tuples of order higher than two). It must be noted that their substantive application rests on non-repeatable events (such as first exit from education) which simplifies the treatment enormously (indeed, I would be tempted to apply the measures they construct in a conventional logistic regression context, given the application to discriminating Italian from Austrian patterns; such use would also depend on the non-repeatable nature of the events).

3.2.3 THE ADVANTAGE OF OMA AS EXPLORATORY

I like OMA, but I do not think it is extremely powerful. Sequence analysis may develop in time into a stochastic technique, with which we can formally test hypotheses and make inferences, but that will not be for a (statistical) generation or two! For the moment its advantage is that it opens longitudinal data to exploration, which is really useful. Again, I will illustrate this from my own experience. In Halpin and Chan (1998) we analysed large samples of class careers (over 1,000 careers at a time), as

V-VI Supervisory and skilled manual
VIIa Semi- and unskilled manual
VIIb agricultural labour
X awaiting entry to labour market (not an egg category)

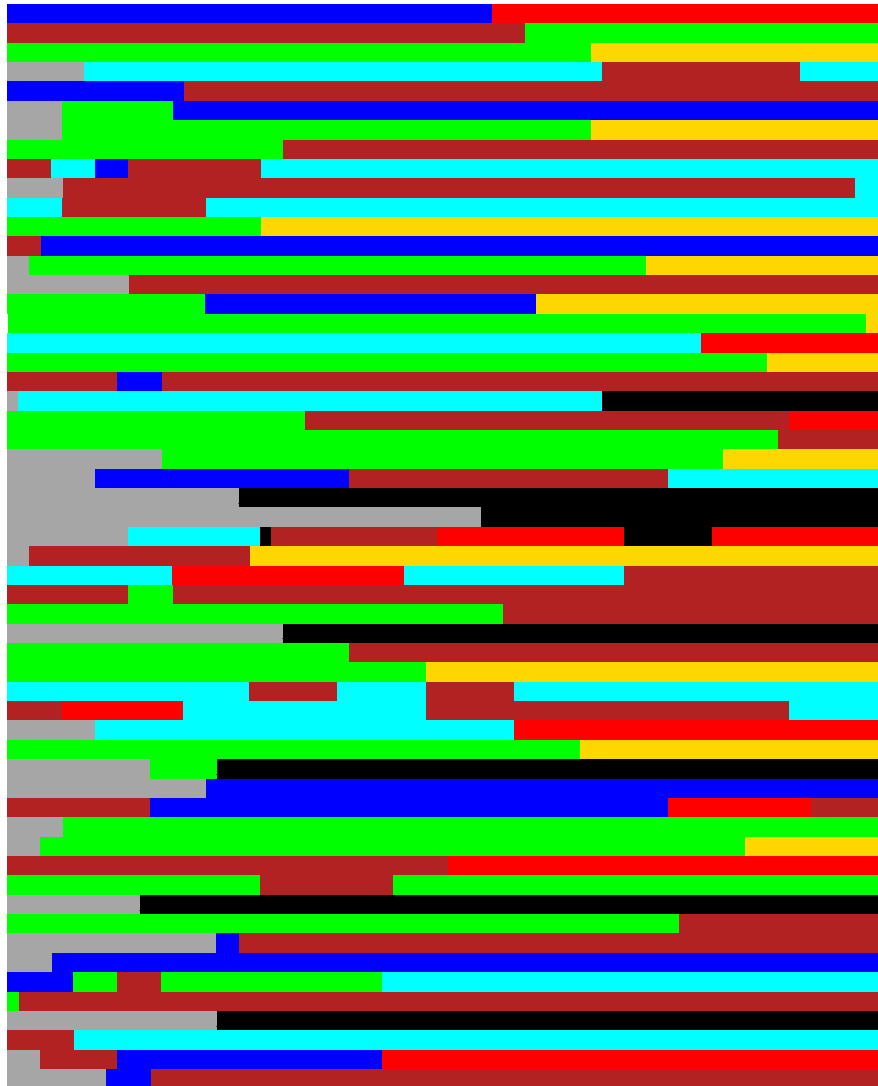


Figure 2: A random sample of class careers

80 unit strings covering age 15 to age 35. Given an interest in the careers as wholes, how do we get an overview of the data set? We can calculate summaries (cumulated durations in different states, origin–destination pairs, and so on) but this is laborious and incomplete. Browsing tens of hundreds of sequences is not terribly easy, even when colour coded, as in Figure 2 which is a random sample of 56 of these sequences.

It may seem like a trivial aspect of the OMA and cluster analysis, but I found it extremely helpful to view the sequences once they had been clustered. Figure 3 shows an entire data set of over 1,000 80-unit careers (class careers from age 15 to age 35 from the Irish Mobility Study (Jackson, 1974)), with sequences grouped by the cluster analysis. Given the approximate nature of clustering, the clusters are more or less disparate (where they contain dissimilar subclusters) but the imposition of order on the large data set makes it far easier to browse and to take in the range of sequences that exist. We can literally see how SA and clustering have worked, noting for instance how the duration sensitiveness has sometimes separated similar careers on the basis of the timing of transitions, how certain states are “absorbing”, how others are most often preceded by particular waiting states, how most people experience few (two or three) transitions, and so on.

Coupling this with the more conventional use of the clusters as an empirical typology (easily “validated” by this visual exercise, for what that is worth) this allows us to get a handle on a sort of data that are hard to digest otherwise.

However, this is clearly in the exploratory domain. If you have hypotheses you want to test, or specific questions you wish to ask, use stochastic techniques. EHA or loglinear modelling or cross-sectional time-series approaches or something else can be applied if you know what you are looking for.

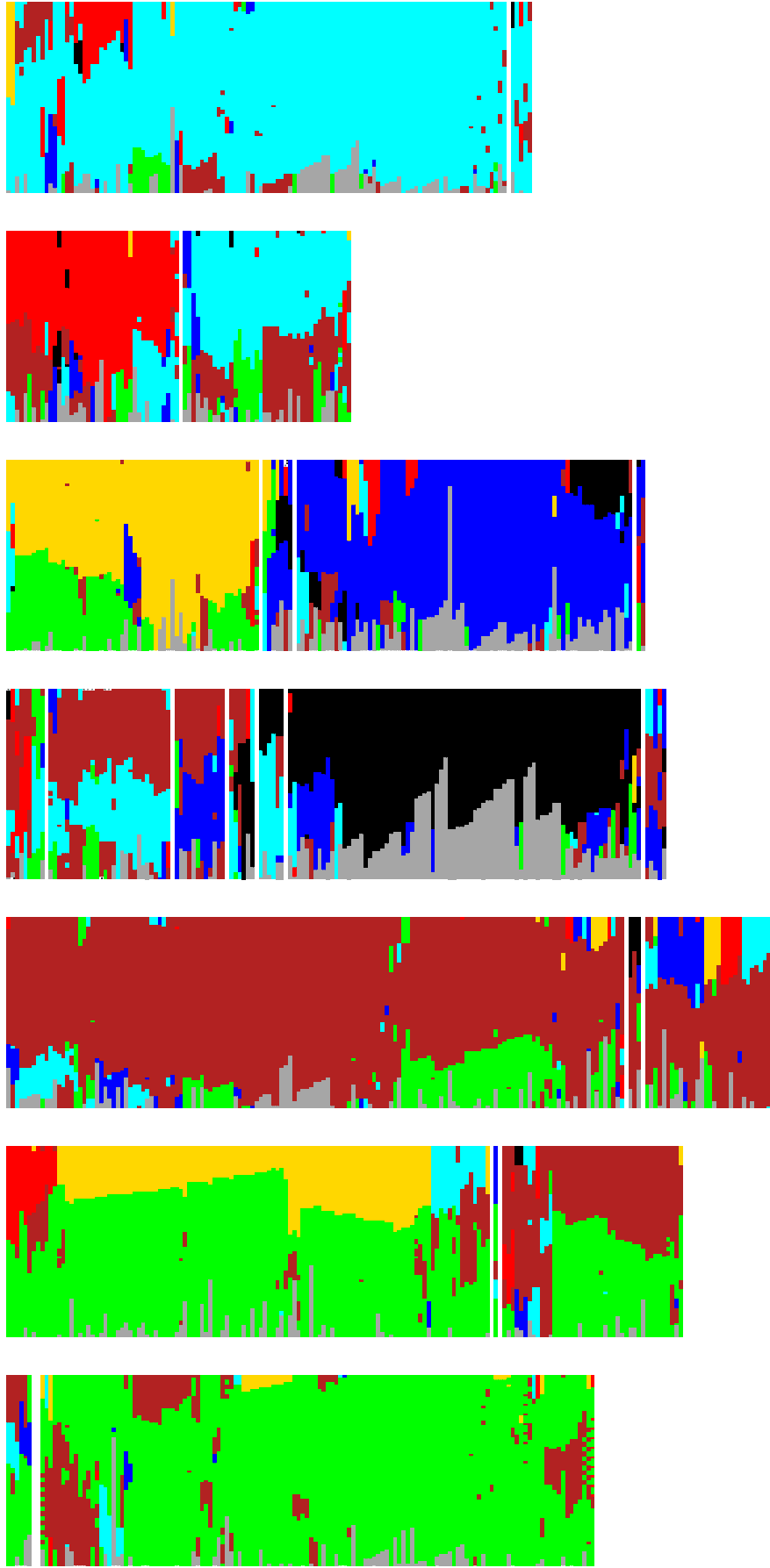


Figure 3: Clustered class careers from the Irish Mobility Study, 1973. Each horizontal line represents an 80-unit summary of the class career from age 15 to age 35, with in total over 1,000 such careers represented. The grouping results from a cluster analysis conducted after optimal matching. The 8-valued class variable is the same as in Figure 2.

4 Exploration is not a crime

And finally, a note on exploration. Much sociological writing on epistemology focusses on theory testing; reading Goldthorpe (2001) recently I find I largely agree with his statements, for instance, about the compatibility of quantitative sociology with ethnography, but what struck me was the absence of consideration of the value of exploration – the things we find out without looking for them or when we do not know what we ought to look for. Given Goldthorpe’s overriding concern with the ‘integration of research and theory’ in that book, it is perhaps not surprising that he does not deal with the incidental accretion of knowledge. In contrast I feel it should be stressed that empirical research activities often teach us far more than strictly allowed by Popper (and that to value such learning is not the same as endorsing empiricism!).

Merton (1949) made some wonderful points about the value of serendipitous discovery, the salient fact we observe by accident while looking for something else, which goes on to stimulate theoretical development. The accidental acquisition of knowledge makes a great difference, and for that we need to have open eyes and to be receptive. Sequence analysis in its current incarnations helps us do that with life history data, and for that alone it is worth while.

References

- Abbott, A. (1988) Transcending general linear reality, *Sociological Theory*, 6, 169–186.
- Abbott, A. (1995) A comment on “Measuring the agreement between sequences”, *Sociological Methods and Research*, 24(2), 232–243.
- Abbott, A. (2000) Reply to Levine and Wu, *Sociological Methods and Research*, 29(1), 65–76.
- Abbott, A. and Barman, E. (1997) Sequence comparison via alignment and Gibbs sampling: A formal analysis of the emergence of the modern sociological article, *Sociological Methodology*.
- Abbott, A. and Forrest, J. (1986) Optimal matching methods for historical sequences, *Journal of Interdisciplinary History*, XVI(3), 471–494.
- Abbott, A. and Tsay, A. (2000) Sequence analysis and optional matching methods in sociology, *Sociological Methods and Research*, 29(1), 3–33.
- Allison, P. (1984) *Event History Analysis*. Beverly Hills, Sage.
- Billari, F. (2002) Sequence analysis in demographic research and applications, in G. Hullen (ed), *Living Arrangements and Households: Methods and Results of Demographic Projections*, Bundesinstitut für Bevölkerungsforschung, Wiesbaden.

- Billari, F., Fürnkranz, J. and Prskawetz, A. (2000) Timing, sequencing and quantum of life course events: a machine learning approach, *Working paper WP 2000-010*, Max Planck Institute for Demographic Research, Rostock.
- Blossfeld, H.-P., Timm, A. and Dasko, F. (forthcoming) Who marries whom in Germany?, in Blossfeld and Timm (forthcoming).
- Blossfeld, H.-P. and Timm, A. (eds) (forthcoming) *Who Marries Whom: Educational Systems as Marriage Markets in Modern Societies*. Kluwer.
- Bradley, D. W. and Bradley, R. A. (1983) Application of sequence comparison to the study of bird songs, in D. Sankoff and J. B. Kruskal (eds), *Time Warps, String Edits and Macromolecules*, Addison-Wesley, Reading, MA, chapter 6.
- Buchmann, M. and Sacchi, S. (1995) Mehrdimensionale Klassifikation beruflicher Verlaufsdaten: Eine Anwendung auf Berufslaufbahnen zweier Schweizer Geburtskohorten, *Kölner Zeitschrift für Soziologie und Sozialpsychologie*, **47**(3), 413–442.
- Chan, T. W. and Halpin, B. (forthcoming) Who marries whom in Great Britain?, *Who Marries Whom? Educational Systems as Marriage Markets in Modern Societies*, Kluwer.
- Coleman, J. S. (1962) Analysis of social structures and simulation of social processes with electronic computers, in H. S. Guetzkow (ed), *Simulation in Social Science: Readings*, Prentice-Hall, Englewood Cliffs, NJ, 61–79.
- Degenne, A., Lebeaux, M.-O. and Mounier, L. (1996) Typologies d'itinéraires comme instrument d'analyse du marché du travail. Troisièmes journées d'études Céreq-Cérétim-Lasmas IdL, Rennes, 23–24 May 1996.
- Dijkstra, W. and Taris, T. (1995) Measuring the agreement between sequences, *Sociological Methods and Research*, **24**(2), 214–231.
- Elzinga, C. (2003) Sequence similarity: A non-aligning technique, *Sociological Methods and Research*. forthcoming.
- Goldthorpe, J. H. (2001) Sociological ethnography today: Problems and possibilities, *On Sociology: Numbers, Narratives and the Integration of Research and Theory*, Oxford University Press, Oxford.
- Greene, W. (2003) *Econometric Analysis*, 5th edn. New Jersey, Prentice Hall. incomplete reference.
- Hagenaars, J. A. (1990) *Categorical Longitudinal Data: Log-Linear Panel, Trend and Cohort Analysis*. Newbury Park, Sage.
- Halpin, B. and Chan, T. W. (1998) Class careers as sequences: An optimal matching analysis of work-life histories, *European Sociological Review*, **14**(2).
- Halpin, B. and Chan, T. W. (2002) Educational homogamy in the UK and

- Ireland: Trends and patterns. Unpublished paper, <http://wivenhoe.staff8.ul.ie/~brendan/papers/edhom2.pdf>.
- Jackson, J. A. (1974) Determinants of occupational mobility in Northern Ireland and the Irish Republic. Computer files, ESRC Data Archive.
- Levine, J. H. (2000) But what have you done for us lately? Commentary on Abbott and Tsay, *Sociological Methods and Research*, **29**(1), 34–40.
- Mare, R. D. (1991) Five decades of educational assortative mating, *American Sociological Review*, **56**, 15–32.
- McVicar, D. and Anyadike-Danes, M. (2002) Predicting successful and unsuccessful transitions from school to work using sequence methods, *Journal of the Royal Statistical Society (Series A)*, **165**, 317–334.
- Merton, R. K. (1949) The bearing of empirical research on sociological theory, *Social Theory and Social Structure*, Free Press, Glencoe.
- Raferty, A. (2001) Statistics in sociology, 1950–2000: A vignette, *Sociological Methodology*.
- Rogoff Ramsøy, N. (1975) On social stratification in a temporal framework. Mimeo, paper presented at OECD Seminar on Education, Inequality and Life Chances, Paris, January 1975.
- Rohwer, G. (1998) *Transition Data Analysis*. Computer program.
- Scherer, S. (2001) Early career patterns: A comparison of Great Britain and West Germany, *European Sociological Review*, **17**(2), 119–144.
- Sørensen, A. B. and Tuma, N. B. (1981) Labor market structures and job mobility, *Research in Stratification and Mobility*, **1**.
- Tuma, N. B. and Hannan, M. T. (1984) *Social Dynamics: Models and Methods*. Orlando, Academic Press.
- Wu, L. L. (2000) Some comments on “Sequence analysis and optimal matching methods in sociology: Review and prospect”, *Sociological Methods and Research*, **29**(1), 41–64.